PB80-223720

Guidelines for the Evaluation of Air Quality Data

(U.S.) Environmental Protection Agency
Research Triangle Park, NC

Feb 74

U.S. Department of Commerce
National Technical Information Service
NTIS.

BIBLIOGRAPHIC INFORMATION

PB80-223720

Guidelines for the Evaluation of Air Quality Data.

Feb 74

PERFORMER: Environmental Protection Agency, Research Triangle Park, NC. Office of Air Quality Planning and Standards.
OAQPS-1.2-015

The purpose of this guideline doucment is to present the basic elements of air quality data analysis that are essential in preparing reports describing the air quality status of a given region. With this aim in mind, emphasis has been placed upon describing both the conventions and the methodology to be employed with minimum discussion of the associated statistical theory. Much of the material that is presented has been treated before but for the sake of completeness, is reiterated in this document with appropriate references indicated.

KEYWORDS: *Air pollution, *Guidelines.

Available from the National Technical Information Service, Springfield, Va. 22161

PRICE CODE: PC A03/MF A01

PB80-223720

FEBRUARY 1974

GUIDELINE SERIES

OAQPS NO. 1.2-015

GUIDELINES FOR THE EVALUATION OF AIR QUALITY DATA

U. S. ENVIRONMENTAL PROTECTION AGENCY
OFFICE OF AIR QUALITY PLANNING AND STANDARDS
MONITORING AND DATA ANALYSIS DIVISION
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

TABLE OF CONTENTS

		PAGE
PREFACE		i
1.	INTRODUCTION	1
2.	BASIC CONVENTIONS FOR HANDLING AIR QUALITY DATA	2
2.1.	Significant Figures	3
2.2.	Minimum Detectable Limit	3
3.	CHARACTERISTIC PATTERNS OF AIR QUALITY DATA	5
3.1.	Seasonal Patterns	7
3.2.	Diurnal Patterns	7
3.3.	Frequency Distribution	10
4.	SUMMARIZING AIR QUALITY DATA	10
4.1.	Indicating Typical Values	13
4.2.	Indicating Maximum Values	15
4.3.	Indicators of Spread	17
5.	MAKING INFERENCES FROM AIR QUALITY DATA	17
5.1.	Inferences About a Particular Site	19
5.2.	Inferences About a Region	22
6.	SOME STATISTICAL TESTS	24
6.1.	Student's T-test	26
6.2.	Non-Parametric Quantile Test	28
7.	BASIC MEANS OF OBTAINING AIR QUALITY DATA	29

PREFACE

The Monitoring and Data Analysis Division of the Office of Air Quality Planning and Standards has prepared this guideline entitled "Guidelines for the Evaluation of Air Quality Data" for use by the Regional Offices of the Environmental Protection Agency. The purpose of the report is to provide guidance information on current air quality data evaluation techniques. Adherence to the guidance presented in the report will, hopefully, ensure mutually compatible ambient air quality data evaluation by all States and Regions. Further, any risks involved in policy decisions concerning National Ambient Air Quality Standards should be minimized. This report will serve on an interim basis until more specific and detailed guidance on this subject is developed.

1. INTRODUCTION

The purpose of this guideline document is to present the basic elements of air quality data analysis that are essential in preparing reports describing the air quality status of a given region. With this aim in mind, emphasis has been placed upon describing both the conventions and the methodology to be employed with minimum discussion of the associated statistical theory. Much of the material that is presented has been treated before but for the sake of completeness, is reiterated in this document with appropriate references indicated.

Since the phrase "air quality data" covers a variety of possible data sets, it is convenient to indicate the exact nature of this phrase as used in this paper. For present purposes, the term "air quality data" refers to a set of observations for a particular pollutant having the following properties:

1. All measurements were made at the same site.
2. Uniform methodology was employed.
3. All measurements have the same averaging time.

It should be noted that the statistical treatments described here for such a data set constitute a minimum effort. There are a variety of more sophisticated techniques available that could be used to extract more information from the data. In general, the degree of effort devoted to data analysis should

be consistent with the value associated with the data. This can be viewed in financial terms as cost of data analysis versus cost of data collection or cost of data analysis versus potential cost of control strategies, etc. In most cases, the extent of the data analysis phase is determined by a subjective judgment of what is appropriate. It should be noted that no matter how extensive the data analysis effort is, the end result can be no better than the original data. This point is particularly important because throughout the following discussions no analysis is made concerning the errors inherent in the measurement method. Therefore, it is essential that the air quality data analyst be aware of the shortcomings in the data and the conclusions that are "statistically significant" be carefully evaluated to determine if they are "really significant."

2. BASIC CONVENTIONS FOR HANDLING AIR QUALITY DATA

Before discussing the analysis of air quality data, it is essential that certain basic conventions be presented for handling the raw data. These conventions are introduced to prevent the air quality summaries from appearing to be more accurate than the data warrants. These conventions have been discussed previously (Nehls and Akland, 1973) and are repeated here since they are the procedures presently employed by EPA in maintaining the National Aerometric Data Bank.

The two topics treated in this section both relate to the relative precision of the raw data with respect to the methodology employed in obtaining the measurement. The first topic concerns the number of significant figures that should be reported while the second deals with values that are below the minimum detectable limit.

2.1. Significant Figures

The number of significant figures that are meaningful for a particular air quality measurement is limited by the methodology employed. To use more significant figures than is warranted by the sensitivity of the analytical procedure adds no real information and can often be misleading. Table 1 presents the suggested reporting accuracy for raw data for various pollutants. While the conventions apply to the raw data it is also useful to specify the accuracy of geometric and annual means. For simplicity, the general convention is that all means be reported to one more significant digit than the raw data.

2.2. Minimum Detectable Limit

Some reported pollutant measurements are below the limit of detection for the analytical procedure. In such cases, the reported number should be viewed as representing a range from zero to the minimum detectable. However, in order to use such data in computing annual summary statistics such as

TABLE 1 - SUGGESTED REPORTING ACCURACY FOR RAW DATA

Pollutant	Number of Decimal Places	
	ug/m^3	ppm
Suspended Particulate Matter	0	--
Benzene Soluble Organic Matter	1	--
Sulfates	1	--
Nitrates	1	--
Ammonium	1	--
Sulfur Dioxide	0	2
Nitrogen Dioxide	0	2
Nitric Oxide	0	2
Carbon Monoxide	1	0
Total Oxidants	0	2
Total Hydrocarbons	1	1
Ozone	0	3
Methane	1	1

geometric means it is convenient to have a convention indicating what value should be substituted for a measurement below the minimum detectable. As a general rule, each value below the minimum detectable is replaced by a value approximately equal to one-half the minimum detectable. Table 2 indicates selected minimum detectable limits used by the National Aerometric Data Bank (NADB) for various analytical methods. A complete listing may be obtained from the National Air Data Branch, EPA, Research Triangle Park, N. C. 27711. The mid-point substitution was selected after examining the statistical distribution of the data (Nehls and Akland, 1973). It should be noted that in comparing data over several years, a standard minimum detectable should be used unless it has changed by an order of magnitude.

In preparing summary statistics, if more than 25% of the observations are less than the minimum detectable no statistics are computed from the data.

3. CHARACTERISTIC PATTERNS OF AIR QUALITY DATA

Before summarizing any data, some thought should be given to the characteristics of the raw data. This is particularly true of air quality data for which strong seasonal and diurnal patterns may effect the interpretation of the data. For example, the maximum hourly oxidant value for a year based on 4,000 observations could have completely different meanings, depending upon whether the observations were made primarily during the winter or the summer. This section presents

TABLE 2

MINIMUM DETECTABLE LIMITS FOR SELECTED MEASUREMENT TECHNIQUES

Pollutant	Collection Method	Analysis Method	Units	Minimum Detectable
Suspended Particulate	Hi-Vol	Gravimetric	ug/m^3	1.0
Nitrate	Hi-Vol	Reduction-Diazo Coupling	ug/m^3	.05
Sulfate	Hi-Vol	Colorimetric	ug/m^3	.5
Carbon Monoxide	Instrumental	Nondispensive Infra-Red	mg/m^3	.575
Sulfur Dioxide	Gas Bubbler	West-Gaeke Sulfamic Acid	ug/m^3	5.0
Total Oxidants	Instrumental	Colorimetric Neutral KI	ug/m^3	19.6

examples of some of these patterns. The analysis of these patterns can frequently be an end in itself since they provide insight into the behavior of the pollutant. An awareness of these patterns also provides a means for screening the data for anomolous values. It should be noted that while the following discussion is general in nature, the characteristic pattern at a given site is a function of local factors such as emissions and meteorology and as a consequence characteristic pattern may be specific to that site or locality.

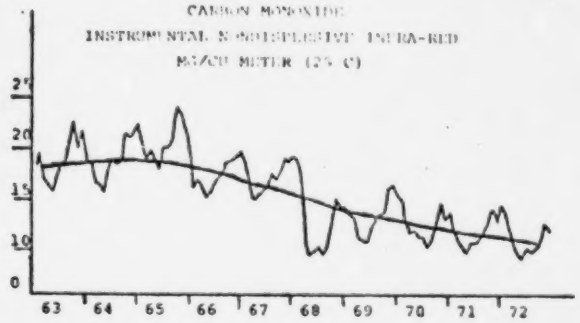
3.1. Seasonal Patterns

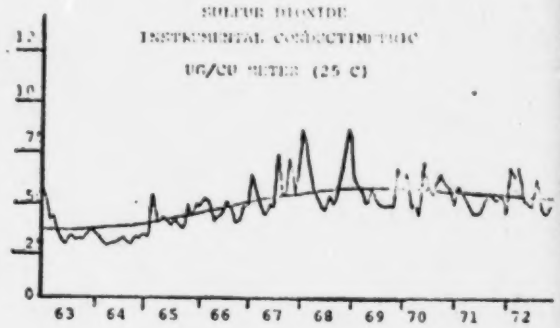
Figure 1 displays graphs of monthly averages for various pollutants at a particular site. Superimposed on these graphs is a smooth curve selected to emphasize the long term trend in the data. Figure 2 displays smoothed curves illustrating the seasonal patterns in the data. The intensity of the seasonal pattern for a particular pollutant may vary from site to site within an area depending upon factors such as proximity to point sources. A knowledge of the seasonality of a pollutant can provide useful information for interpreting the data since it suggests the season in which maximum concentrations would be expected.

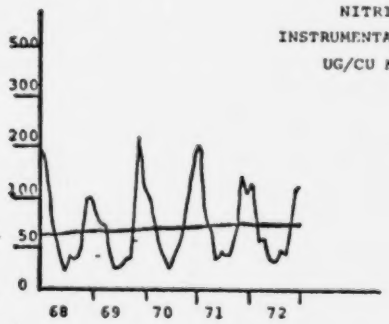
3.2. Diurnal Patterns

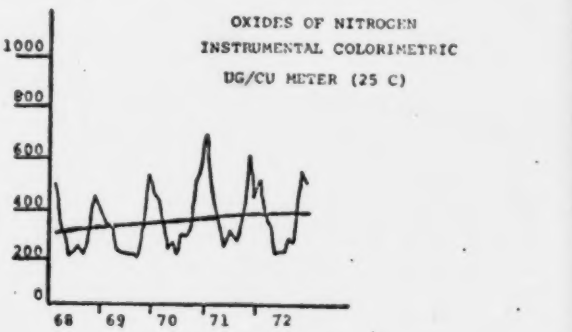
In addition to seasonal patterns some pollutants also have pronounced diurnal patterns. These patterns may be due to factors such as solar radiation, traffic density, etc. which influence pollution levels.

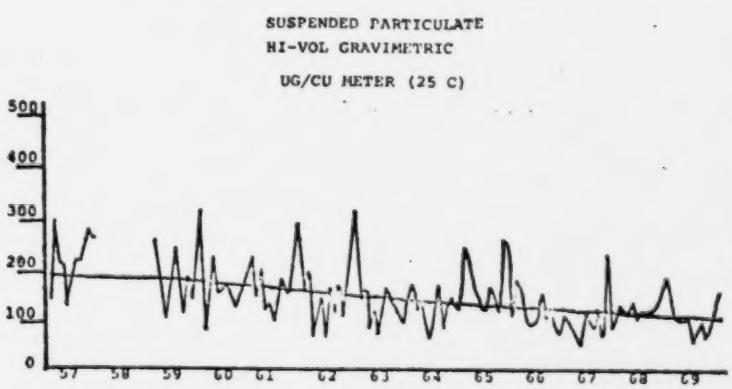
FIGURE CHARTS OF MONTHLY AVERAGES FOR VARIOUS POLLUTANTS AT A PARTICULAR SITE

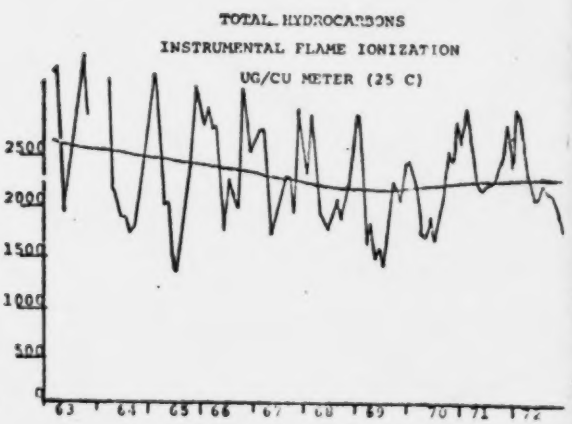


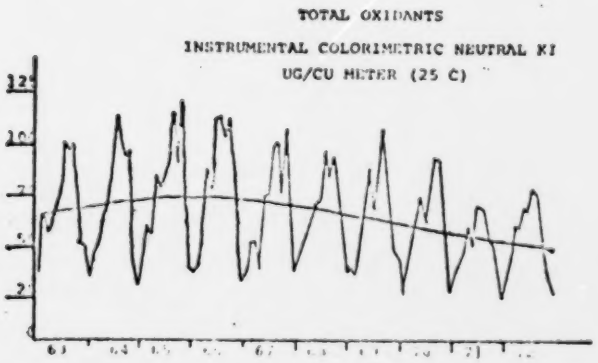


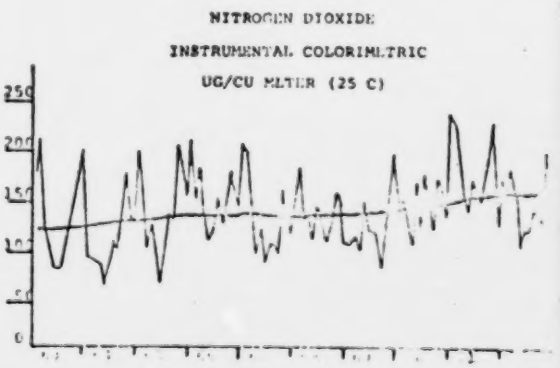




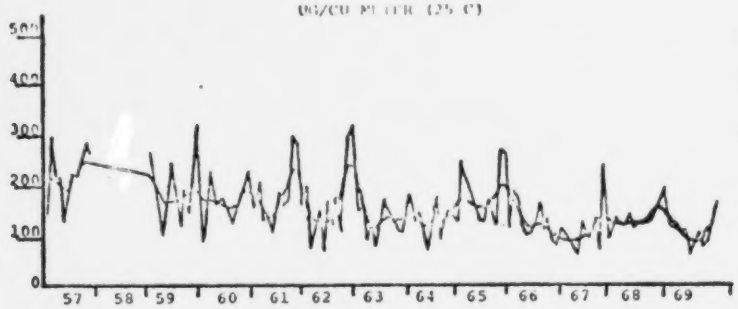


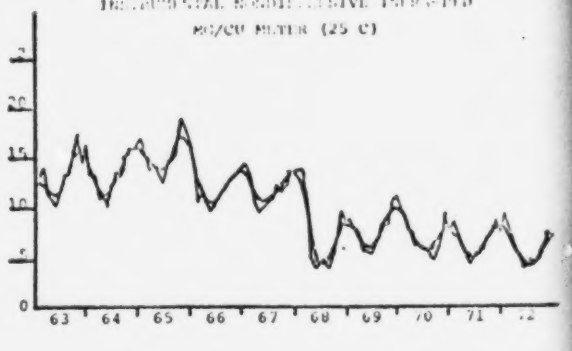


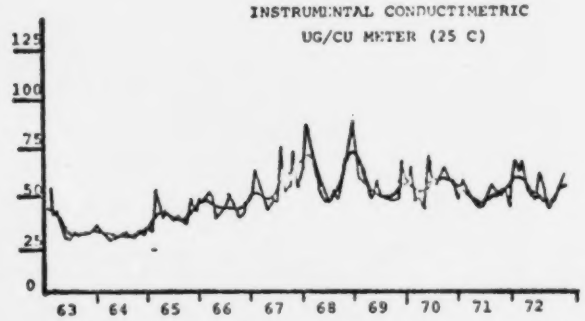


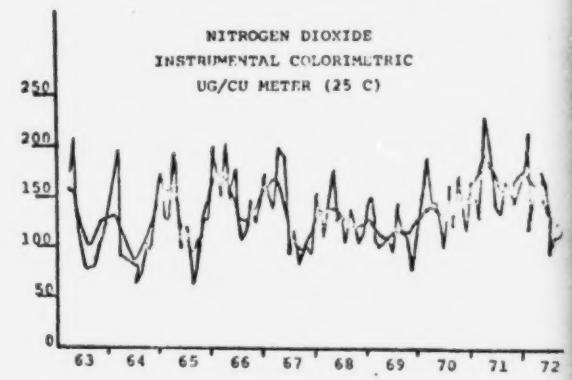


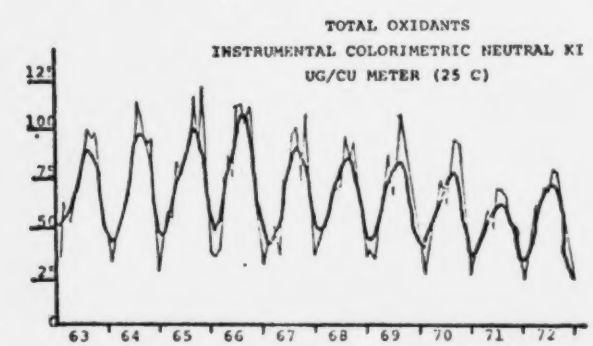
Reproduced from
best available copy.

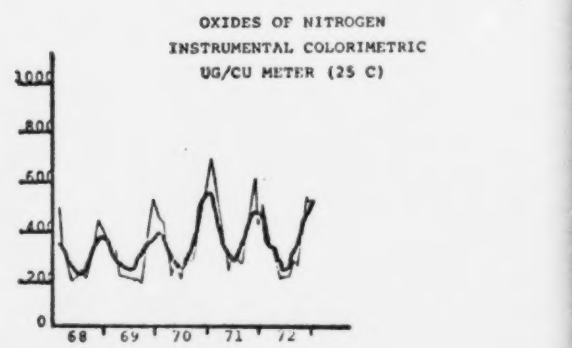


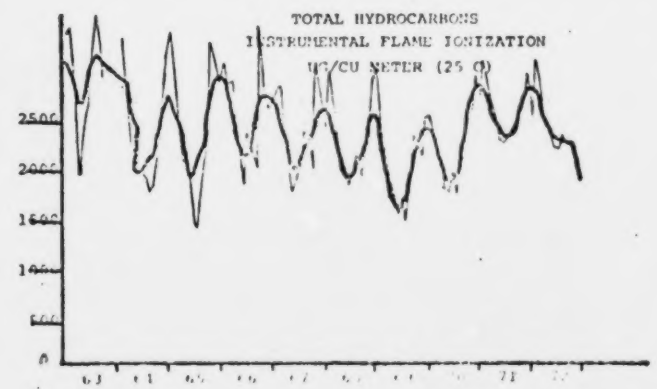


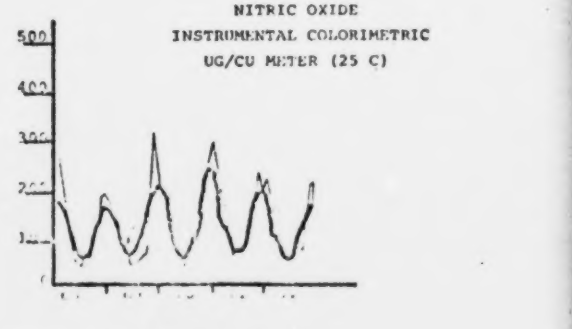












Reproduced from
best available copy.

Table 3 summarizes the 1971 oxidant data for the Downtown Los Angeles sites operated by Los Angeles Air Pollution Control District. The number of times that the national oxidant standard was exceeded is presented by month and hour of the day. The marginal totals indicate both the diurnal pattern and the seasonal pattern.

3.3. Frequency Distributions

One characteristic pattern of air quality data that is particularly important becomes apparent after examining some frequency distributions. Many quantities are assumed to have a symmetric distribution about the average such as the normal distribution. Figure 3 shows the frequency distribution for total suspended particulate data from Philadelphia. It is apparent that this distribution is not symmetric. However, Figure 4 shows the frequency distribution for the logs of this same data. The distribution is more symmetric and may be approximated by a normal curve. Data having this property is said to be log-normally distributed and this is a common assumption regarding air quality data (Larsen, 1971).

4. SUMMARIZING AIR QUALITY DATA

In preparing a summary of air quality data, one of the most important steps is to determine the purpose of the summary. The usual use of these summaries is to indicate typical levels and peak levels. This section discusses some of the basic statistics that can be used for this purpose.

TABLE 3 NUMBER OF HOURS ABOVE OXIDANT STANDARD BY MONTH AND TIME OF DAY (1971 DATA)

DOWNTOWN LOS ANGELES

	M	1	2	3	4	5	6	7	8	9	10	11	N	1	2	3	4	5	6	7	8	9	10	11	TOTAL BY MONTH
JAN													1	2	2	3									8
FEB												1	4	4	4	3									16
MAR										1	1	1	3	3	2	1									12
APR											4	6	8	8	7	7	3	1							44
MAY												3	4	4	3	1	1								16
JUN									1	2	9	9	12	12	11	6	2	1							65
JUL										2	13	19	18	15	11	4		1							83
AUG										2	8	17	16	16	7	3	1								70
SEPT										3	6	10	10	10	6	1									46
OCT											2	7	5	9	6	2									31
NOV													1	1											2
DEC																									0
TOTAL BY HOUR									1	10	43	73	82	84	59	31	7	3							393

FIGURE 3 - FREQUENCY DISTRIBUTION - TSP (PHILADELPHIA-1969)

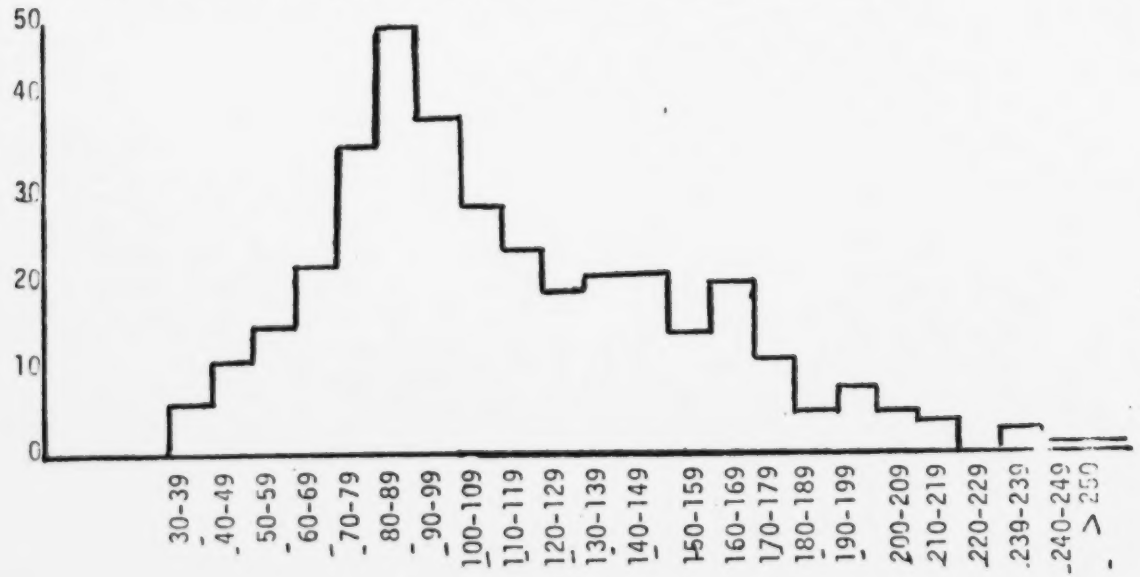
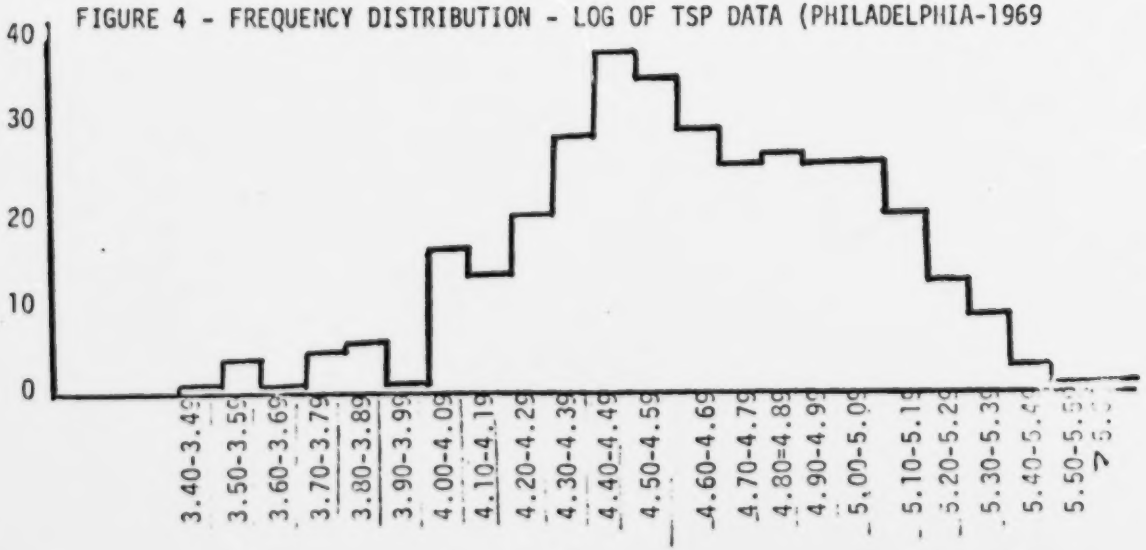


FIGURE 4 - FREQUENCY DISTRIBUTION - LOG OF TSP DATA (PHILADELPHIA-1969



The first two subsections discuss the treatment of typical and peak values. The third discusses the range of the data.

4.1. Indicating Typical Values

This section discusses the arithmetic mean, the median, and the geometric mean as indicators of typical values. The arithmetic mean and the median are frequently used in air pollution studies because of certain properties of the log-normal distribution. In choosing the appropriate statistic, the purpose of the summary must be considered. While all three may indicate typical values, if the purpose of the summary is to compare the data to the National Ambient Air Quality Standards, then the standard suggests the appropriate statistic. A commonly used statistic to indicate typical values is the mode. The mode is the value that occurs most frequently. The use of the mode is not discussed here since it is frequently of little value in summarizing air quality data. For example, the mode for oxidant could be near the minimum detectable due to low values throughout the night.

Arithmetic Mean

Given a set of n observations, say $X_1, X_2, \ldots, X_n$, the arithmetic mean is simply $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

When the term "average" is used the arithmetic mean is usually what is meant.

Median

The median is the middle value of the data. That is if the data is ranked in order of magnitude so that $X_1 \leq X_2 \ldots \leq X_n$ then the median is $X_{\frac{n+1}{2}}$ if n is odd, and $\left(\frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}\right)$ if n is even.

The median is a convenient statistic that is not influenced as much as the arithmetic mean by changes in the extremely high or low values of the distribution.

Geometric Mean

Given a set of n observations, say X_1, X_2 ... X_n, the geometric mean is $g = (X_1 \cdot X_2 \ldots X_n)^{1/n}$.

Since this probably is the least intuitive of the statistics presented, it is worthwhile to discuss it in more detail.

If distribution is symmetric, such as the normal distribution, then the expected value of the arithmetic mean and median are identical. However, for a log-normally distributed variable, it is the expected value of the geometric mean that approximates the expected value of the median. Therefore, since some air pollutants have a distribution that is approximately log-normal, the geometric mean became used as a convenient method of summarizing the data and for total suspended particulate, the annual standards are expressed as geometric means.

As an alternate computational formula, it should be noted that

$$\log g = \frac{1}{n} \sum_{i=1}^{M} \log x_i \text{ or } g = \text{EXP}\left\{\frac{1}{n} \sum_{i=1}^{n} \log x_i\right\}.$$

4.2. Indicating Maximum Values

As in the previous section, the purpose of the summary is a critical factor in determining the appropriate statistic. Maximum values may be indicated by listing the maximum and/or the second highest value. The second highest value is important since compliance with the short-term air quality standards is determined by this value. However, there are other statistics that are useful for indicating maximum values. The principle difficulty with using the second highest value is that it does not allow for differences in sample sizes. For example, if two monitoring devices are side by side and one monitors every day of the year while the other monitors only every sixth day, it would be expected that the second high value for the every day device would be higher than the every sixth day device even though both monitored the same air. Table 4 illustrates how the second high value may vary depending upon different sampling frequencies based upon total suspended particulate data from a Philadelphia site that sampled daily.

To allow for this dependence upon sample size, various percentiles are sometimes used to indicate maximum values. For example, the 99th percentile might be used for hourly data while the 90th might be appropriate for daily measurements.

TABLE 4 MAXIMUM AND SECOND HIGH VALUES (PHILADELPHIA-1969)
FOR VARIOUS SAMPLING SCHEMES

Sampling Schedule	Observations	Maximum	Second Highest
Everyday	365	325	244
Every Sixth Day	61	219	215
"	61	195	171
"	61	244	238
"	61	215	211
"	61	325	234
"	60	239	205
Every Fifteenth Day	25	205	176
"	25	325	207
"	25	239	191
"	25	219	196
"	25	234	165
"	24	201	198
"	24	215	211
"	24	195	183
"	24	188	173
"	24	195	169
"	24	160	154
"	24	244	199
"	24	215	201
"	24	179	171
"	24	238	205

By using a percentile value, allowance is made for varying sampling frequencies from site to site and year to year. Table 5 indicates the 90th percentile for the sampling schedules used in Table 4.

4.3. Indicators of Spread

In addition to an indication of typical and peak values, it is also desirable to have a measure of how variable the data is. Did it fluctuate widely or were all values fairly uniform? The customary statistics for this purpose are either the arithmetic standard deviation or the geometric standard deviation. Ranges or percentiles could also be used depending upon the desired use of the summary but they are not discussed. The basic formulas for the arithmetic and the geometric standard deviations are given below.

Let $X_1, X_2, \ldots, X_n$ be a set of n observations. Then the arithmetic standard deviation is:

$$s = \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{1/2} \quad \text{where } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and the geometric standard deviation is

$$s_g = \text{EXP}\left[\frac{1}{n} \sum_{i=1}^{n} (\ln X_i - \ln g)^2\right]^{1/2}$$

where g is the geometric mean.

5. MAKING INFERENCES FROM AIR QUALITY DATA

Once the air quality data has been summarized, it is in a convenient form to be examined so that conclusions can be made regarding air quality. At this point the data is either

TABLE 5 GEOMETRIC MEANS, MEDIANS, AND 90TH PERCENTILE VALUES FOR SAMPLING DATA OF TABLE 4

Sampling Schedule	Observations	Geometric Mean	Median	90th Percentile
Everyday	365	102.6	97	171
Every Sixth Day	61	99.8	105	162
"	61	95.2	93	155
"	61	113.6	113	188
"	61	107.2	101	177
"	61	106.4	105	171
"	60	94.7	94	158
Every Fifteenth Day	25	100.2	111	175
"	25	114.6	121	178
"	25	125.0	130	189
"	25	104.9	96	192
"	25	100.8	105	148
"	24	99.8	90	190
"	24	104.4	98	177
"	24	102.4	99	171
"	24	92.1	95	143
"	24	100.8	96	162
"	24	92.0	88	140
"	24	104.6	97	186
"	24	107.2	109	173
"	24	94.1	94	162
"	24	99.6	98	165

extremely useful or extremely dangerous depending upon the quality of the summary. This section discusses these inferences to illustrate the potential dangers that can result from inadequate summaries. For convenience, the discussion is divided into two parts. The first deals with inferences about a particular site while the second deals with inferences about a region.

5.1. Inferences About a Particular Site

This section discusses inferences that can be made about a given site from one year's data for a particular pollutant. Since any conclusions based upon the data can be no better than the data itself, the most important part of the summary is to decide if the data gives adequate annual coverage. This relates directly to the previous discussion of characteristic patterns. If an annual average is to be computed from the data, then it is essential that all portions of the year be represented equally. An examination of the seasonality that exists for certain pollutants shows why this is essential. As a convenient rule, it may be assumed that if each calendar quarter contains at least 20% of the total observations then the sample is adequately balanced. If this is not the case, then a more appropriate way to determine the annual average is to use a weighted mean calculated as follows:

(1) determine the average for each quarter and

(2) compute the average of these four quarterly averages.

While the previous constraint applies to the seasonal balance of the sample, it is also essential to have a restriction on the minimum number of observations that are required to compute an annual mean. Such constraints are employed in the National Aerometric Data Bank system (Nehls and Akland, 1973) and to maintain uniformity, they are repeated here. For continuous

measurements at least 75% of the total possible observations should be present before summary statistics are calculated. The exact requirements are given in Table 6. For intermittent sampling data, there must be at least five observations per quarter and if one month has no observations the remaining two months in that quarter must both have at least two observations. While these conventions are used in general, it is of course possible to modify them for certain applications. For the most part the general intention of these restrictions is to ensure that the observations are sufficiently representative of the entire year to calculate an annual mean. For peak value statistics such as the number of times a certain value is exceeded the constraint is not essential in showing violations. For example, two hourly oxidant values in excess of the standard is sufficient to show non-compliance even if there were no other observations that year. Nevertheless, to assess the extent of the problem, data sufficient to meet the requirements for determining a mean would be advantageous although for seasonal pollutants it could suffice to summarize only particular quarters or months.

In discussing the inferences that can be made from a given sample, it is worth observing that while the annual mean can be either under- or over-estimated the maximum and the second high values can only be underestimated assuming no instrumental error. For example, if a simple hypergeometric probability

TABLE 6 SUMMARY CRITERIA FOR CONTINUOUS MEASUREMENTS

Time Interval	Minimum Number of Observations
3-hour running average	3 consecutive hourly observations
8-hour running average	6 hourly observations
24-hour	18 hourly observations
Monthly	21 daily averages
Quarterly	3 consecutive monthly averages
Yearly	9 monthly averages with at least two monthly averages per quarter

model is assumed, Table 7 shows the probability of detecting violations of the short-term standard as a function of sampling frequency. From this table it may be seen that if samples are taken every sixth day the probability of detecting two excursions above the standard is less than 50% unless the site actually exceeds the standard 10 days per year. This illustrates the weaknesses associated with determining maximum values on the basis of intermittent sampling.

Two possible solutions to this problem are (1) to intensify sampling schedules or (2) to use mathematical equations to extrapolate from the data to predict maximum values. At the present time, there is no convenient predictive formula that can be applied on a general basis to give sufficiently accurate maximum values. As a guide, the predictive formula developed by Larsen (1971) based on the log-normal distribution may be used to determine the possible magnitude of the underestimation due to intermittent sampling. However, this empirical model assumes log-normality and independence and should not be used to determine compliance with the standards since its predictive accuracy has not been fully documented.

5.2. Inferences About a Region

Once conclusions have been made for each site in a region the next step is to draw conclusions concerning the region. If any one of the sites exceeds the NAAQS then the region is not in compliance. It should also be pointed out that the worst

TABLE 7 PROBABILITY OF SELECTING TWO OR MORE DAYS WHEN SITE IS ABOVE STANDARD

Actual no. of excursions	Sampling Frequency - Days per year 61/365	122/365	183/365
2	.03	.11	.25
4	.13	.41	.69
6	.26	.65	.89
8	.40	.81	.96
10	.52	.90	.99
12	.62	.95	.99
14	.71	.97	.99
16	.78	.98	.99
18	.83	.99	.99
20	.87	.99	.99
22	.91	.99	.99
24	.93	.99	.99
26	.95	.99	.99

site in the region may still underestimate the magnitude of the air pollution problem. The only way in which a site may overestimate the air pollution problem is if it is not representative of the air to which receptors are exposed. There are guideline documents discussing this subject. While it is relatively easy to compare the air quality in a region with the NAAQS it is not so easy to compare one region with another. For example, one region may choose to concentrate most of its monitoring efforts at sites having high pollution potential while another region may have numerous sites monitoring background levels. Therefore, extreme caution should be used if such comparisons must be made and particular attention should be given to the placement of monitoring sites.

6. Some Statistical Tests

When making inferences from air quality data it is frequently necessary to have some objective means to make judgments. This is the point at which statistical inference becomes useful. The previous treatment has used statistics merely for descriptive purposes in order to conveniently summarize the data. The purpose of statistical inference is to objectively substantiate generalizations made from the data. For this reason, two basic statistical tests are discussed.

While these statistical tests are relatively straight forward, a certain degree of caution is required regarding the underlying assumptions that determine their validity. Since one of these assumptions is particularly important in applications dealing

with air quality data, it will be discussed in detail.

In statistics, it is commonly assumed that the data to be analyzed is a random sample of all the data and that the measurements are independent. While this may be approximately true for intermittent data collected on a sampling scheme comparable to that employed by the NASN, it may not be true for all samples. For the most part, these statistical assumptions are merely a mathematical formulation of common sense ideas. Certainly, if data were only collected on Sundays, it would not be expected that the average of these numbers is truely representative of the annual average. Sampling schedules that only monitor certain days of the week result in non-random samples and their degree of usefulness is inherently limited. The problem of independence is somewhat more subtle. For example, successive hourly oxidant measurements are not independent. While the concept of statistical independence may be clearly defined in mathematical terms, it is possible to present an intuitive notion of what it entails. Two numbers may be thought of as being independent if knowing one of the numbers does not help in guessing what the other number is. The classical example of this is rolling dice in which knowing what number occurred on one die does not improve a guess of what number occurred on the other. With this in mind, it is apparent that knowing one hourly oxidant value helps in guessing what the next hourly value will be. It

should be noted that it is not necessary that it make the guess a certainty-only that it improve the chances of guessing correctly.

With the ideas of randomness and independence in mind, it is possible to present two statistical techniques that are generally useful in practice. The first test is commonly known as student's t-test and is useful for examining the mean. The second test is the non-parametric quantile test and despite the rather elegant name it is a convenient test for the median and other percentiles and is very easy to use.

6.1. Student's t-test

The Student's t-test is a commonly used statistical test for data that may be assumed to be normally distributed. As mentioned earlier, air pollution is frequently assumed to be log-normally distributed so that the t-test may be employed to examine the logarithms of the data. The application of this technique to determine confidence intervals for annual goemetric means has been discussed by Hunt (1972) and is briefly treated here. This present discussion examines construction of a confidence interval for an annual mean. Extensions to comparisons of two means may also be performed but are not treated here since the approach is almost identical and can be found in basic statistical texts. More general tests concerning trends at a site are examined in the guideline document for trend analysis.[5]

The basic application is that a set of data from an intermittent monitoring device has been obtained. This data has been used to determine the annual geometric mean. Since this data re-

presents only a fraction of the total number of days in the year, the question arises as to how close the mean of the data is to the actual annual mean. The statistical technique employed for this purpose is the confidence interval so that a probability statement may be made regarding the range of the true annual mean.

To calculate a 95% confidence interval for the geometric mean, the interval is first constructed for the arithmetic mean of the logarithms. To do this, the following calculations are necessary:

Let $\bar{x}_{log} = \frac{1}{n} \sum_{i=1} \log x_i$, where n is the sample size

Let $s_{log} = \left[\frac{\sum_{i=1}^{n} (\log x_i - \bar{x}_{log})^2}{n-1} \right]^{1/2}$

Let $d = t_{1-\alpha/2} \frac{s_{log}}{\sqrt{n}} (1-\frac{n}{N})^{1/2}$ where $t_{1-\alpha/2}$ is obtained from a table for Student's t-test where $1-\alpha$ is the confidence level and N is the possible number of samples, e.g. 365 for daily samples.

Then the lower and upper confidence intervals for the geometric mean, denoted as L and U respectively, are given by

$L = EXP(\bar{x}_{log} - d)$

and $U = EXP(\bar{x}_{log} + d)$.

It should be noted that in the above formulas the finite correction factor, $(1-\frac{n}{N})$, was used since it is assumed that the

population size is finite rather than infinite. For example, in considering daily measurements it is assumed that the population size is 365, i.e. the total number of days in the year.

6.2. Non-Parametric Quantile Test

In discussing the t-test it was pointed out that it is necessary to assume that the logarithms of the air pollution measurements are normally distributed. In some cases, it may not be desirable to make this assumption. For example, an examination of the data may show that such an assumption is unwarranted. For such cases, non-parametric statistical tests are appropriate since they do not require any assumptions regarding the form of the underlying distribution. Moreover, non-parametric tests are frequently quite easy to employ since many of the calculations are relatively simple. A variety of non-parametric tests are available. A more detailed description of the test discussed here is available in the text by Conover (1971).

Quantile is a more general term than percentile. For the present discussion, the test is used to examine the median but it may also be applied to any percentiles or quantiles. It is also assumed that there are more than 20 observations since this is generally true for air quality problems and reduces the need for tables.

Let $x_1, x_2, \ldots, x_n$ be a sample of air quality measurements and suppose it is desired to test if the annual median is greater than a specific value, say s.

Then it is only necessary to calculate the following two values:

T = the number of sample values less than or equal to s

and $t = pn + w_\alpha \sqrt{np(1-p)}$, where n is the sample size p is the quantile value and w_α is the α quantile of a standard normal random variable.

For tests at the .05 level w_α is - 1.645.

For tests concerning the median the quantile value is .5 so the above formula becomes

$$t = .5n - 1.645 \sqrt{.25n}$$

$$= .5n - .822 \sqrt{n} .$$

If T is less than t then the conclusion may be stated that "the median is greater than s" and that the result was obtained by employing the quantile test "at the 5% level."

7. Basic Means of Obtaining Air Quality Data

One station continuously monitoring oxidant can produce 8,760 observations. Therefore, considerable caution should be exercised when requesting air quality data since there is a considerable risk of being inundated with unnecessary numbers. Usually when questions arise concerning air quality, the answer may be given in terms of summary statistics and it is not necessary to review the raw data. Certain basic sources include the various

periodic reports from State and local agencies as well as EPA's reports on the NASN and CAMP monitoring efforts. Overview reports with extensive appendices such as The National Air Monitoring Program: Air Quality and Emission Trends Annual Report, are also available.

The National Aerometric Data Bank provides many summary files that may be accessed by time sharing terminals. In addition, the NADB provides printouts containing general information that may be easily looked up with no need to access the computer. Table 8 lists frequent questions and a readily available source.

TABLE 8 NADB OUTPUT FOR COMMON QUESTIONS ON AIR QUALITY

Question	Source
What data is available nationwide for a particular pollutant?	Inventory by pollutant
What data is available for a particular geographical region?	Inventory by site
What was maximum value at a site (annual)?	Any inventory
What was mean value at a site (annual)?	Any inventory - if valid year
How many observations (annual)?	Any inventory
Status of a site with respect to NAAQS?	"Standards Program"
Frequency Distribution	Time Sharing Option (TSO)
Quarterly or monthly data	Time Sharing Option
Raw data	Time Sharing Option
Description of the site such as UTM coordinates, county, operating agency, etc.	Site File

REFERENCES

1. Conover, W. J., "Practical Nonparametric Statistics," John Wiley and Sons, Inc., New York, 1971.

2. Hunt, W. F., Jr., "The Precision Associated with the Sampling Frequency of Log-Normally Distributed Air Pollutant Measurements," Journal of the Air Pollution Control Association, Volume 22, No. 4, September 1972.

3. Larsen, R. I., "A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards," AP-89, November 1971.

4. Nehls, G. J. and G. G. Akland, "Procedures for Handling Aerometric Data," Journal of the Air Pollution Control Association, Volume 23, No. 3, March 1973.

5. "Guidelines for the Evaluation of Air Quality Trends, Interim Guidelines," U. S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, N. C., OAQPS No. 1.2-014, December 1973.